big data and cognitive computing

MDPI

Article

Beneficial Artificial Intelligence Coordination by Means of a Value Sensitive Design Approach

Steven Umbrello

Institute for Ethics and Emerging Technologies, Via San Massimo 4, Turin 10123, Italy; Steve@ieet.org

Received: 18 December 2018; Accepted: 2 January 2019; Published: 6 January 2019

check for updates

Abstract: This paper argues that the Value Sensitive Design (VSD) methodology provides a principled approach to embedding common values into AI systems both early and throughout the design process. To do so, it draws on an important case study: the evidence and final report of the UK Select Committee on Artificial Intelligence. This empirical investigation shows that the different and often disparate stakeholder groups that are implicated in AI design and use share some common values that can be used to further strengthen design coordination efforts. VSD is shown to be both able to distill these common values as well as provide a framework for stakeholder coordination.

Keywords: value sensitive design; VSD; design for values; safe for design; AI; ethics

1. Introduction

Value Sensitive Design (VSD) is a design methodology that begins with the premise that technologies are value-laden and that human values are continually implemented both during and after the design of a technology [1,2]. The 'sensitivity' of VSD is to the values that are held by the multitude of stakeholders that are both directly and indirectly enrolled during technological design whether they be engineers, CEOs and/or the relevant publics. This paper aims to argue for the VSD approach as a potentially suitable methodology for artificial intelligence coordination between the often-disparate publics, governmental bodies, and industry. In evaluating the applicability of VSD to AI coordination, this paper eschews any in-depth discussion of superintelligence or AI risk scenarios. In doing so, the aim of this paper is to lay out arguments for the adoption of VSD that can have an immediate impact on existing AI systems and on the systems of the near future. The value of this immediacy is taken for granted given the urgency proposed by the abundant AI risk research.

VSD exists among various other safe-by-design methodologies within the field of responsible research and innovation (RRI) and itself comes in various forms depending on the domain of applications [3–6]. It is largely agreed in the design literature, spanning back to the inception of technologies studies that technology is not value-neutral, but rather that values are consistently implicated in design [7,8]. Artificial intelligence, like robotics, nanotechnology, and information and communication technologies (ICTs), among others, is a sociotechnical structure that implicates not only the physical, or digital entity itself, but also the infrastructures, people and politics that it emerges from and into [9–14]. Not only this, but sociotechnical systems function only in accordance with the boundaries of this social context, they require actors and institutions that constrain and direct developmental pathways towards certain avenues rather than others [15,16]. The actors and infrastructures that allow a sociotechnical system to emerge naturally implicate values with questions such as: which funding bodies are permitted to distribute monies? How are research avenues chosen and who judges what is an acceptable research stream? How are opportunity-cost decisions made and under what criteria are some paths chosen rather than others? Because each of these questions is naturally implicated in design and because each of them implicates values, values in design must be

considered more carefully, not only of the technologies in question themselves but also the institutions and social infrastructures that enroll these values.

VSD provides such a way to evaluate the values that are implicated both on technical and social dimensions as has been demonstrated in its application for other socio-technical systems [17–20]. Dignum et al. (2016) and Oosterlaken (2015) both explore the potential application of applying the VSD framework to socio-technical energy systems, whereas Umbrello and De Bellis (2018) explore more explicitly the potential boons that a VSD approach can bear on the technical development of intelligent agents (IA). Umbrello and De Bellis (2018) provide a theoretical basis for which moral values of stakeholders could be designed into the technical systems of IAs and provides means for adjudicating moral overload [21], however, they do not give any real account of how VSD could ameliorate the gap between various, often conflicting stakeholders. Dignum et al. (2016), however, provide a valuable analysis of various groups such as the federal government, non-governmental organizations (NGOs) and commercial organizations with regards to the surveying and extraction of shale gas in the Netherlands. In evaluating the policy documents of these different stakeholders, the authors were able to infer and distill a set of root values. However, although both Dignum et al. (2016) and Oosterlaken (2015) provide useful studies, they do not give any empirical case for the application of VSD to existing sociotechnical systems. Mouter, Geest, and Doorn (2018) argue that because the Dutch government scuttled the exploitation of the shale gas in the Netherlands, there was no way for Dignum et al. (2016) to elicit the explicit design considerations that can be used for a thorough VSD analysis [22].

To the best of my knowledge, this paper is the first to evaluate the merits of the VSD framework for AI coordination per se. Prior literature on VSD has focused on its methodology [8,23], its application to existent technologies [24,25], its philosophical underpinnings [26,27] and even to the reduction of future AI risk [20]. These studies provide useful information regarding both VSD and AI but do not provide any tangible analysis of the issues of coordination, nor to those that are particular to AI. This paper's application of the VSD approach as a means to ameliorate the often-disparate stakeholders that are implicated in the development and use of AI technologies is particularly unique. It is similarly the intent of this paper to spark further research on some of the issues regarding how VSD can be used to coordinate stakeholders of other technological innovations that converge with AI, such as nanotechnology and biotechnology.

To successfully tackle this argument, this article is organized into the following sections (see graphical abstract): the first section will lay out the methodological framework of the VSD approach as well as how it has been applied to other technological innovations. In doing so, one can begin to conceptualize both the strengths and potential drawbacks of the VSD approach as it can be formulated for application to AI systems. The second section will draw upon the work done in §1 by beginning to sketch multiple pathways for potential AI coordination by formulating specific examples of coordination between various AI stakeholders by drawing on a specific case study that implicates a variety of stakeholders. In doing so, this paper builds on the previous work done by Umbrello and De Bellis (2018) which explores how the VSD approach can be used to design intelligent agents (IAs) specifically. While that paper explored the technicalities of IA design, this paper investigates the stakeholders themselves to better form pathways for coordination. The final section of this paper sketches broader theoretical implications that these conclusions may have and points to potential future research avenues.

2. Material and Methods

Emerging from the domain of human-computer interaction (HCI) and ICT, VSD has since developed into a largely adopted design approach to incorporate human values (and perhaps even non-human) values during both the early and latter design phases of technologies [23,28]. Since its inceptions in the early 1990s, VSD has been adopted as a proposed framework for the design of identity technologies [25], energy technologies such as wind turbines [19,24], robotics and autonomous

agents such as care robots, autonomous vehicles, and AI in the medical field [20,29–32], information and communication technologies such as sensors and communicative computer software [33–38], health technologies such as ambulatory therapeutic assistance systems and seizure detectors [39–42], and nanotechnology both in its advanced and contemporary forms [43–45]. VSD is described by its founders Batya Friedman et al., as a tripartite framework consisting of conceptual, empirical and technical investigations [23].

Conceptual investigations are characterized as philosophical evaluations of determining who the stakeholders are, determining the values that are identified, what values should be chosen, as well as how conflicts between values are to be resolved. Next, empirical investigations use various surveying methods such as observations and interviews, as well as other explorative tools to determine if the values distilled in conceptual investigations can be successfully embedded into a certain technological design [1]. The third investigation, technical investigations, is characterized by two steps: the first determines how the technology under question constrains or supports humans values whereas the second avenue determines how the distilled values of the conceptual investigations can be sufficiently embedded in the technological design [46]. Although empirical and technical investigations are complimentary and akin to one another, the difference between the two is not insignificant. Empirical investigations focus primarily on stakeholders who are affected, either directly or indirectly by the technological design whereas technical investigations investigate the technology per se.

VSD is often chosen over competing theories because its emphasis is not only on the conceptualization of the values that are embedded, or aim to be embedded in a design, but because it requires adding an empirical and technical analysis to evaluate the role of systems and institutions that affect design as well as how stakeholder groups form a co-constitutive role in a technologies safe-adoption [47]. The importance here for AI stakeholders is that VSD provides a principled way of engaging with different stakeholder groups, giving a way for their values and perceptions of AI to be formulated into a root set of instrumental values that can then be brought directly into the design process. Lastly, the framework may tally benefits to the design practice by determining moral overload a priori, establishing understanding within and between stakeholder groups regarding potentially emerging value-conflicts. Moral overload in the design literature refers to when elicited stakeholders provide conflicting, yet still important values for technological design [21]. What VSD does not do however is provide a clear way of *actually* embedding values into a design. Its aim is to highlight the root values at play by stakeholders and to determine if the technology in question supports or constrains those values [48], however, formulated the notion of a 'value hierarchy' (see Figure 1) that allows the moral values of stakeholders to be more easily conceptualized as functional design requirements [48].

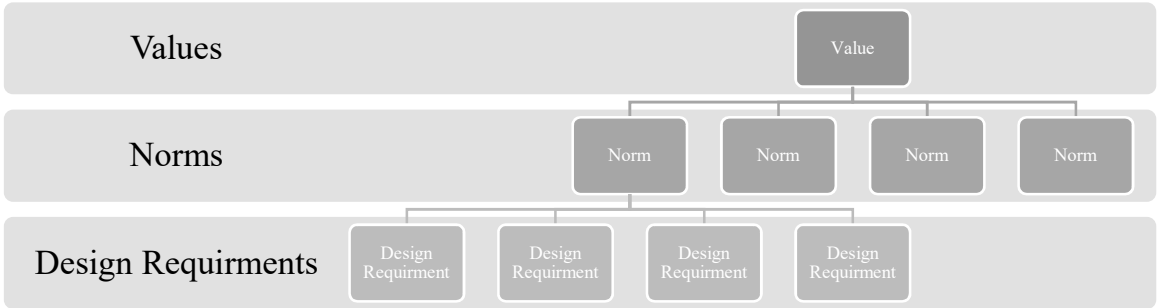

Figure 1. Top-Down Values hierarchy (Source: [48]).

What this paper does then, in order to better conceptualize how different stakeholders relevant to AI conceptualize values, is use Van de Poel's value hierarchy as the main tool to construct a set of root values that can aid to bridge the cooperative design gap. A top-down hierarchy of values such as Figure 1 consists of three distinct ranks, the top rank (Values) is objective. It is objective in the

sense that the root values distilled are not sensitive to context [27] or culture. For example, [26] argues against this very notion, arguing for both intersubjectivity as a means by which to reconceptualize VSD as well as the reformulation of VSD away from moral law theories towards an imaginative theory of morality that is more in line with modern neuroscience. The proceeding rank consists of norms, which inhere as every form of imperative or constraint on action, these differ from the root values of the higher-order rank of values because norms are sensitive to context and situation. The lowest rank aims to formalize the higher-order rank of norms as functional design requirements. In doing so, the norms aim to be translated into an applied practice that can then be introduced into the design flow [48–50].

However, the hierarchy need not flow in the top-down direction as the original formulators of VSD originally conceptualized; it can similarly move from the bottom upwards. It begins naturally with a particular set of existing design requirements that are then used to distill a common set of root values. The following section of this paper employs this dual-directional analysis (best conceptualized by Figure 2) to better find a path of cooperation between AI stakeholders.

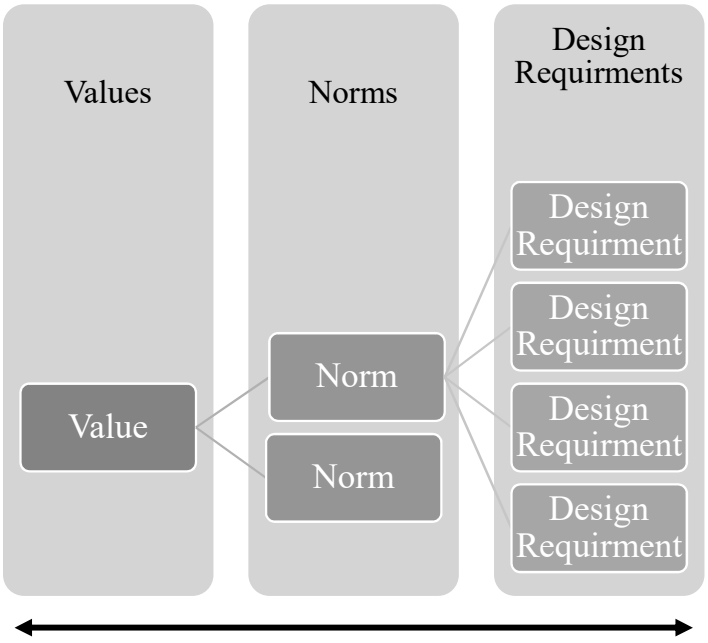

Figure 2. Bi-Directional Values Hierarchy.

The purpose of this paper is to determine the suitability of the VSD approach to the coordination of various stakeholders involved and implicated in beneficial AI [51] research and development. In doing so, it draws upon one potentially controversial case, that of the appointment of the UK Select Committee on Artificial Intelligence. This particular case has been selected over other controversial cases because (1) its ad hoc nature gives it a discrete time-specificity and ease by which the case can be analyzed, and (2) the case did and continues to garner media scrutiny. Because of both (1) and (2), coupled with the potential societal influence that the committee can have as a result; the ability to source relevant material and literature is straightforward and accessible.

In the second report of the 2016–17 session of the House of Lords Liaison Committee—an advisory group to the House which advises, oversees, and reviews the resources needed for the selection and coordination of select committees and ad hoc committees—advised for the formation of four ad hoc committees of which the subject of one was solely to focus on artificial intelligence [52]. These ad hoc committees, selected in the 2016–17 session, were established as year-long seats, which were then to report their findings in time for the 2017–18 session in March 2018.

Acknowledging the impacts of continued technological advances, proposals for the establishment of an ad hoc select committee on artificial intelligence were forwarded to focus on the economic, social

and ethical issues implicated by the design and use of artificial intelligence systems. Because it is a topic of specific interest that does not fall within the purview of the expertise of any existing committee (i.e., it is the first of its kind), the establishment of a topic-specific committee was decided upon. More specifically, the ad hoc committee was envisioned to evaluate the impact of AI on the following topics, taking into account both the arguments of the 'techno-optimists' and the 'techno-pessimists':

- Pace of technological change
    - Relationship between developments in artificial intelligence and productivity growth;
    - Creation of new jobs;
    - Sectors and occupations most at threat from automation.
- Economic and social issues
    - The role of government in the event of widespread job displacement;
    - Further education and training, for both children and adults;
    - Unemployment support, including the case for a universal basic income;
    - Government funding for artificial intelligence-related research and development.
- Ethical issues
    - The government's role in monitoring the safety and fairness of artificial intelligence;
    - Transparency around the use of 'big data';
    - Privacy rights of individuals;
    - General principles for the development and application of artificial intelligence. (Source: [52])

From 29 June 2017, when the appointments of the Select Committee on AI were established, the members met in three closed sessions over the course of the month. The following meeting was their visit to DeepMind on 13 September 2017. The following months consisted of a combination of both closed private sessions as well as public evidence sessions of which transcripts of the panels are fully accessible online [53]. After several closed sessions between January and March 2018, the Select Committee's final report was published on 16 April 2018 and later publicly debated in government on 19 November 2018.

The final report concluded that the UK is well positioned to be a global leader in AI research and development. Properly designed and implemented, the report considered the UK to be in a unique position to address social, economic and ethical issues that existed and that may arise with the design and implementation of AI system and take advantage of the economic and social benefits that they are predicted to usher. Similarly, the report acknowledges the value-ladenness of technologies, their socialtechnicity, and the past issues of prejudice being designed into technological systems; the resolution was taking care in the early design phases to ensure an equitable design process.

Finally, the report argues for more transparent access to data and the enrollment of stakeholders into the decision-making processes of industry and governmental bodies directly responsible for the design of AI. Presently, discussions of practical steps to bridge cooperative gaps are taking place to apply the recommendations of the committee's report.

As already outlined, the VSD approach was originally construed as an anticipatory design framework that envisioned a technological design in isolation from the socialtechnicity that it was to emerge in. However, the already widespread use of AI systems makes a purely ex-ante approach impotent, and for this reason, both the top-down and bottom-up rankings are required. These permit adjustments and modifications as new information makes itself known [54].

To this end, in this section, I uncover some of the most pertinent values of ethical importance within the context of this case. Typically, as per the original instantiations of the VSD approach, the vast

body of philosophical and sociological literature is levied to better distill a set of core values. Friedman et al., along with [20] provide a strong point of departure within the realm of both HCI and AI regarding potentially relevant values such as safety, privacy, accountability, and sustainability [20,55]. The remainder of the list of values (Table 1) is drawn from the various written and oral transcripts that eventually formed the collated evidence volumes that were gathered by the Select Committee [56]. As such, what follows is an *empirical investigation* as per the VSD approach given by the committee themselves engaged in the conceptual investigations of determining the ethical values implicated in AI.

The written comprehensive evidence volume consists of 223 separate reports by policy experts, academics, NGOs, think tanks, governmental bodies, and industry leaders [56]. This categorization employed in this paper to separate the different evidence reports and testimonies is taken directly from the reports themselves which are explicit in their affiliation and category. Similarly, the oral evidence volume consists of 57 separate oral testimonies by similar groups and individuals [57]. Likewise, the government response to the House of Lords Artificial Intelligence Select Committee's report provides a clear perspective on how the UK aims to address the report's findings [58]. What should be noted here is that the represented sample size garnered by the reports (and by the committee's search) do not reflect a full sample size of stakeholders affected (or can be affected indirectly) by AI technologies. The values distilled are those projected by the 'experts' appointed by the committee to draw reports. Because of this, this paper, as well as the case study as a whole, represent an initial sketch of how conceptual investigations can be undertaken, and are an illustration of the further work that needs to be done in order to draw a representative stakeholder group that accounts for population from the considered area, in what concerns its structure: age, gender, occupation, educational level, family size.

The bi-directional approach to distilling values and design requirements is of particular use when investigating these documents given that their eclectic sources, ranging from not only those listed but also those with both philosophical and engineering backgrounds. The ability to use both approaches to come to a similar set of values and design requirements permits a more thorough approach to determining's a common list of values, even if it only serves as a starting point for collaborative actions between the relevant stakeholders implicated in the government's proceedings.

3. Results

To this end, the list of values in Table 1 is the result of a prolonged distillation of the bi-directional method. Each of the 223 separate written evidence reports, as well as the transcripts of the 57 oral witness testimonies, were read for both an explicit account of what needed to be construed as a design requirement (i.e., a value) whereas norms and technical design requirements were contextualized into values. What resulted is a major overlap of a series of 12 values ranging in support. Transparency was shown to be the most widely supported, overlapping with 146 different reports. The majority of the evidence reports employed the term transparency, while others preferred interpretability or 'explainability', sometimes interchangeably. The final report opted for the use of 'intelligibility' to refer to the broader issue. Similarly, intelligibility can be approached in two distinct ways: (1) technical transparency and (2) Explainability. Similarly, control and data privacy came in both second and third, respectively, in terms of support by the different evidence reports (see Figure 3 for the rank-order distribution).

Prescriptions for technical *transparency* to permit users and designers to understand how and why the decisions made by AI systems were taken was one of the most identified top-down values. Technical recommendations, like the ability for both users and designers to access a system's source code, were the primary norms identified, however, that, per se, does not entail transparency for why certain decisions were chosen over others, nor does it show the data input that leads to those decisions. Similarly, transparency was argued to be a value that is contingent on the stakeholder group in question, as well as the purpose of the AI system in question. For example, Professor Chris Reed, Professor of Electronic Commerce Law, Queen Mary University of London, argued that:

There is an important distinction to be made between ex-ante transparency, where the decision-making process can be explained in advance of the AI being used, and ex-post transparency, where the decision-making process is not known in advance but can be discovered by testing the AI's performance in the same circumstances. Any law mandating transparency needs to make it clear which kind of transparency is required [59].

Table 1. The 12 values supported throughout the collected evidence volumes with the number of unique reports that explicitly supported each value or provided a design requirement or norm that could be distilled into a value.

Values	Academics/ Scholars/Universities	NGOs/Think Tanks/Non-Profits	Governmental Bodies	Industry/For Profit
Data Privacy	9	5	5	14
Accessibility	4	5	3	7
Responsibility	41	15	4	18
Accountability	35	10	10	25
Transparency	62	27	13	44
Explainability	5	4	2	4
Efficiency	19	16	5	19
Consent	35	25	5	19
Inclusivity	7	4	5	7
Diversity	26	14	5	24
Security	44	33	9	35
Control	65	35	7	34

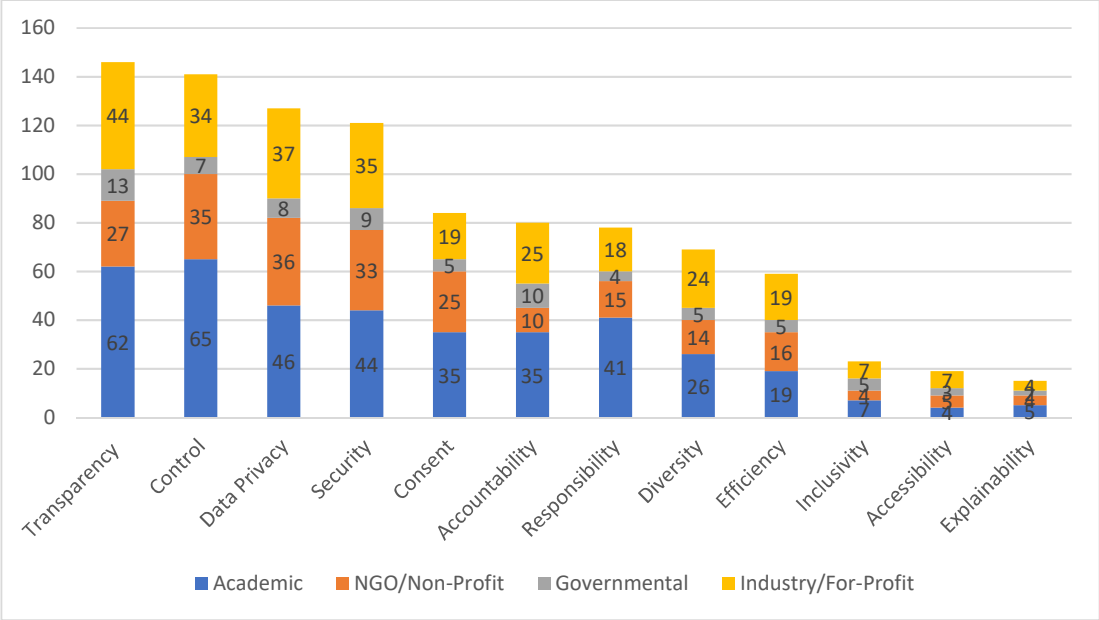

Figure 3. Rank-Order Distribution of Values. Numerical values represent individual report.

Certain constraints on ex-ante transparency thus could be warranted because absolute transparency prior to an AI development could severely curtain AI development and innovations. Nonetheless, sacrifice to innovation in favor of transparency was universally affirmed by the reports where fundamental human rights were at stake.

Diversity and inclusivity, on the other hand, were values that were identified through the bottom-up approach, usually in relation to a more explicit value and how that value can be strengthened or realized through design requirements. The value of transparency, for example, can help to determine what inputs are being fed into a system and determine if those inputs and the subsequent decisions are

impartial, inclusive and diverse. These two values, in particular, were not identified in the top-down approach and were relegated exclusive to design requirements that supported more explicit norms and values.

4. Discussion

So far, this paper has looked at how a specific case study has engaged in conceptual investigations on AI design and development to determine the human values that are important to different stakeholders. Values were identified both from the top-down and bottom-up methods. What follows in this section is a cursory look at how VSD can be further harmonized with the existent and ongoing work in AI to further bridge disparate stakeholder groups.

Transparency, control, and privacy arose in this study as the most explicit values expressed, while values such as diversity, inclusivity, and accessibility were expressed as bottom-up design requirements or norms that were related to securing one of those three values. Because of this, those values, particularly transparency, is used to discuss how the VSD approach could be used to further embed that value into AI design.

In evaluating the content that discussed transparency either explicitly or as a design requirement, the concerns that were mostly expressed were that ex-post technical-approaches to AI systems' transparency is difficult, if not impossible. However, there are nonetheless cases where such transparency is imperative, even if they come at the cost of "power and accuracy" [59]. To this end, transparency can be affirmed through the design requirement of technical explainability, in which ex-ante approaches to systems development require AIs to continually explain the logic and inputs used to arrive at their decisions [60]. The adoption of the VSD approach during preliminary stages of AI development thus can help to mitigate the difficulties of ex-post black boxes and help to determine the level of stakeholder tolerance between competing values such as transparency and privacy. For this reason, the inclusion of foundation norms such as "determining the diversity and inclusivity of data sets" helps to strengthen higher rank-ordered values such as transparency. The inclusion of these norms throughout the design process provides both a path for the formalization of new design requirements, as well as a way to reformulate values in less-obvious ways.

Additionally, the values distilled in both this study, as well as in the collated report should not discount, nor be prioritized over those of continued conceptual investigations by designers. The investigations of values as a purely conceptual, a priori practice aids designers to deliberate on values that may not emerge in stakeholder elicitations. Although the design of AI systems with the explicit values of stakeholders may increase system adoption and acceptance, the values that can emerge through the principled conceptual investigations that VSD formalizes is also of importance. Similarly, given the socio-technicity of AI, stakeholders may often overlook how infrastructures, technical standards, the values of designers, and other social systems constitute and shape the values that are implicated in technological development. Similarly, delimiting who the stakeholders are and adequately selecting a representative group to elicit values is difficult, hence making conceptual investigations an important step along with empirical and technical investigations. In doing so, when designers elicit stakeholder involvement, they can then reflect on the values of conceptual investigations to continually adapt them to the changing technical and empirical input.

Although VSD does not offer the ideal solution for bridging stakeholder groups and solidifying their coordination in the design of AI, it does nonetheless present the fundamentals for (1) determining common values across stakeholder groups through both norms and design requirements (and vice versa) and (2) makes value conflicts functionally apparent and addressable thus (3) permitting both ex ante and ex post interventions to take place that account for a wide variety of stakeholder values. Having a formalized approach like this, with clear stages and delineations, allows designers to design AI systems in a principled way that reduces the likelihood of biased or uninformed decisions. A step that can bet taken by committees and similar groups such as the UK Select Committee on AI is to acknowledge a common set of values amongst the select stakeholders, extend those conceptual and

empirical investigations to other stakeholder groups that were perhaps not considered during the initial conceptual investigations and determine if there is any overlap. Similarly, those values can then be used to determine design requirements that can express those values at technical level in design.

5. Conclusions

The purpose of this paper was to explore the potential applicability of the VSD methodology to the development and fostering of cooperation and collaboration between various stakeholder communities in the design and development of AI systems. Through the application of empirical investigations as outlined in the VSD framework, this paper explored the implicated human values that may be relevant to the design of AI systems. It concluded that, in the case of the UK Select Committee on AI, that a common value hierarchy could be distilled from disparate stakeholder groups and from different mediums of translation (i.e., reports, testimonies, and newspapers). The bi-directional approach to the value-hierarchy was shown to be the best way to distill both values and design requirements given that different mediums offered different ways of arriving at either one (policy reports vs. news reports). Transparency, for example, was always shown through the top-down approach whereas values such as diversity and inclusivity were only through the bottom-up approach. An important observation of this study is that transparency is an important, yet multi-faceted and often difficult, value to incorporate into design, requiring ex-ante interventions at the design stages to increase transparency via technical explainability.

The findings of this paper have the potential to allow both stakeholders and engineers to better conceptualize the values of different groups that may reduce AI recalcitrance and increase stakeholder *inclusivity* and *accessibility*. In doing so, the design process for the multitude of AI systems can be strengthened both from the early design phases and throughout their development through continued stakeholder dialogue.

It is acknowledged that both this paper and VSD have their limitations. The investigations carried out in this particular case study are both socially and culturally situated, and thus limited. Similarly, the values explored by VSD are considered universal rather than socially or culturally relative [26]. Likewise, VSD affirms strong anthropocentrism in its value investigations whereas an abundance of literature from both cultural anthropology and philosophical ecology have shown that the values of nonhuman actors (and perhaps eventually AGI/ASI) are always already implicated in human actions in the Anthropocene [61,62]. This study has shown from where initial steps can be taken towards the design of beneficial AI, but further research studies should not only work from the initial premises of this paper but explore the viability of both non-anthropocentric values as well as the flexibility of the underlying assumptions of VSD's conceptual investigations. Although some recent work has begun these investigations [26–28], it has yet to be adopted as common practice within the design scholarship and requires further argumentation if it is to be so.

Additionally, VSD can be limited in many cases by constraints on the relevant literature to undertake conceptual investigations. Similarly, restricted access to relevant stakeholder groups, diversity, and inclusivity of the members of those groups and the ability to resolve the moral overload of value conflicts in a clear and principled way all limit the VSD methodology. This paper, for example, is not only limited in these ways but it also focuses primarily on empirical investigations and disregards the technical investigations that are critical to VSD.

Nonetheless, what this study has shown is that VSD can be applied both ex-ant and ex-post facto to sociotechnical systems that already exist. What is needed are both research and policy measures that can determine the actual impact of the adoption of VSD as a general framework for design. What VSD aims to do, and this paper should have shown more explicitly, is that through a thorough investigation of various sources and stakeholders, various design requirements can be translated into a common set of held values and that explicit values can also be translated into design requirements. Similarly, the work that has gone into this study to better facilitate the hierarchy of values from these various mediums shows that that VSD methodology with a bi-directional hierarchy approach requires

a substantial time investment to ensure that important values or design requirements are not passed over. Whether this is true for various cultures and social contexts is yet to be seen and can only be done with its wider adoption, if and when that happens. That being said, continued VSD research should similarly look at the situations in which the produced studies emerge to better determine weakness within both the studies themselves and the VSD framework (i.e., improvements could reduce partiality and cultural bias, and give voice to silenced stakeholders).

Funding: This research received no external funding.

Acknowledgments: I would like to thank the two anonymous peer reviewers who have taken the time to both quickly and thoroughly review this manuscript. Their recommendations have undoubtedly increased the quality of this paper. Any remaining errors are the author's alone. The views in the paper are the authors' alone and not the views of the Institute for Ethics and Emerging Technologies.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Friedman, B.; Kahn, P.H.; Borning, A.; Huldtgren, A. Value Sensitive Design and Information Systems. In *Early Engagement and New Technologies: Opening up the Laboratory*; Doorn, N., Schuurbiers, D., van de Poel, I., Gorman, M.E., Eds.; Springer: Dordrecht, The Netherlands, 2013; pp. 55–95.
2. Van den Hoven, J. The Design Turn in Applied Ethics. In *Designing in Ethics*; van den Hoven, J., Miller, S., Pogge, T., Eds.; Cambridge University Press: Cambridge, UK, 2017; pp. 11–31.
3. Boenink, M. The Multiple Practices of Doing 'Ethics in the Laboratory': A Mid-Level Perspective. In *Ethics on the Laboratory Floor*; Springer: Berlin, Germany, 2013; pp. 57–78.
4. Doorn, N.; Schuurbiers, D.; Van de Poel, I.; Gorman, M.E. *Early Engagement and New Technologies: Opening up the Laboratory*; Springer: Berlin, Germany, 2014; Volume 16.
5. Fisher, E.; O'Rourke, M.; Evans, R.; Kennedy, E.B.; Gorman, M.E.; Seager, T.P. Mapping the Integrative Field: Taking Stock of Socio-Technical Collaborations. *J. Responsib. Innov.* **2015**, *2*, 39–61. [CrossRef]
6. Micheletti, C.; Benetti, F. Safe-by-Design Nanotechnology for Safer Cultural Heritage Restoration. Available online: http://atlasofscience.org/safe-by-design-nanotechnology-for-safer-cultural-heritage-restoration/ (accessed on 15 December 2017).
7. Winner, L. Do Artifacts Have Politics? *Technol. Future* **2003**, *109*, 148–164. [CrossRef]
8. Van den Hoven, J.; Manders-Huits, N. Value-Sensitive Design. In *A Companion to the Philosophy of Technology*; Wiley-Blackwell: Oxford, UK, 2009; pp. 477–480.
9. Bijker, W.E.; Hughes, T.P.; Pinch, T. *The Social Construction of Technological Systems: New Directions in the Sociology and History of Technology*; MIT Press: Cambridge, MA, USA, 1987.
10. Bechtold, U.; Fuchs, D.; Gudowsky, N. Imagining Socio-Technical Futures—Challenges and Opportunities for Technology Assessment. *J. Responsib. Innov.* **2017**, *9460*, 1–15. [CrossRef]
11. Pitt, J.; Diaconescu, A. Interactive Self-Governance and Value-Sensitive Design for Self-Organising Socio-Technical Systems. In Proceedings of the 2016 IEEE 1st International Workshops on Foundations and Applications of Self* Systems (FAS*W), Augsburg, Germany, 12–16 September 2016; pp. 30–35. [CrossRef]
12. Baxter, G.; Sommerville, I. Socio-Technical Systems: From Design Methods to Systems Engineering. *Interact. Comput.* **2011**, *23*, 4–17. [CrossRef]
13. Trist, E. The Evolution of Socio-Technical Systems. *Occas. Pap.* **1981**, *2*, 1981.
14. Crabu, S. Nanomedicine in the Making. Expectations, Scientific Narrations and Materiality. *TECNOSCIENZA Ital. J. Sci. Technol. Stud.* **2014**, *5*, 43–66.
15. Williamson, O.E. Transaction-Cost Economics: The Governance of Contractual Relations. *J. Law Econ.* **1979**, *22*, 233–261. [CrossRef]
16. Wüstenhagen, R.; Wolsink, M.; Bürer, M.J. Social Acceptance of Renewable Energy Innovation: An Introduction to the Concept. *Energy Policy* **2007**, *35*, 2683–2691. [CrossRef]
17. Künneke, R.; Mehos, D.C.; Hillerbrand, R.; Hemmes, K. Understanding Values Embedded in Offshore Wind Energy Systems: Toward a Purposeful Institutional and Technological Design. *Environ. Sci. Policy* **2015**, *53*, 118–129. [CrossRef]

18. Dignum, M.; Correljé, A.; Cuppen, E.; Pesch, U.; Taebi, B. Contested Technologies and Design for Values: The Case of Shale Gas. *Sci. Eng. Ethics* **2016**, *22*, 1171–1191. [CrossRef]
19. Oosterlaken, I. Applying Value Sensitive Design (VSD) to Wind Turbines and Wind Parks: An Exploration. *Sci. Eng. Ethics* **2014**, *21*, 359–379. [CrossRef] [PubMed]
20. Umbrello, S.; De Bellis, A.F. A Value-Sensitive Design Approach to Intelligent Agents. In *Artificial Intelligence Safety and Security*; Yampolskiy, R.V., Ed.; CRC Press: Boca Raton, FL, USA, 2018; pp. 395–410.
21. Van den Hoven, J.; Lokhorst, G.J.; van de Poel, I. Engineering and the Problem of Moral Overload. *Sci. Eng. Ethics* **2012**, *18*, 143–155. [CrossRef] [PubMed]
22. Mouter, N.; de Geest, A.; Doorn, N. A Values-Based Approach to Energy Controversies: Value-Sensitive Design Applied to the Groningen Gas Controversy in the Netherlands. *Energy Policy* **2018**, *122*, 639–648. [CrossRef]
23. Friedman, B.; Hendry, D.G.; Borning, A. A Survey of Value Sensitive Design Methods. *Found. Trends®Hum.–Comput. Interact.* **2017**, *11*, 63–125. [CrossRef]
24. Correljé, A.; Cuppen, E.; Dignum, M.; Pesch, U.; Taebi, B. Responsible Innovation in Energy Projects: Values in the Design of Technologies, Institutions and Stakeholder Interactions 1 (Draft Version for Forthcoming Book) Aad Correljé, Eefje Cuppen, Marloes Dignum, Udo Pesch & Behnam Taebi. In *Responsible Innovation 2*; Koops, B.-J., Oosterlaken, I., Romijn, H., Swierstra, T., van den Hoven, J., Eds.; Springer International Publishing: Cham, Switzerland, 2015; pp. 183–200.
25. Briggs, P.; Thomas, L. An Inclusive, Value Sensitive Design Perspective on Future Identity Technologies. *ACM Trans. Comput. Interact.* **2015**, *22*, 1–28. [CrossRef]
26. Umbrello, S. The Moral Psychology of Value Sensitive Design: The Methodological Issues of Moral Intuitions for Responsible Innovation. *J. Responsib. Innov.* **2018**, *5*, 186–200. [CrossRef]
27. Umbrello, S. Imaginative Value Sensitive Design: How Moral Imagination Exceeds Moral Law Theories in Informing Responsible Innovation. Masters Thesis, University of Edinburgh, Edinburgh, UK, 2018. [CrossRef]
28. Umbrello, S. Safe-(for Whom?)-By-Design: Adopting a Posthumanist Ethics for Technology Design. Masters Thesis, York University, Toronto, ON, Canada, 2018. [CrossRef]
29. van Wynsberghe, A. Designing Robots for Care: Care Centered Value-Sensitive Design. *Sci. Eng. Ethics* **2013**, *19*, 407–433. [CrossRef] [PubMed]
30. Santoni de Sio, F.; van den Hoven, J. Meaningful Human Control over Autonomous Systems: A Philosophical Account. *Front. Robot. AI* **2018**, *5*, 15. [CrossRef]
31. Thornton, S.M.; Lewis, F.E.; Zhang, V.; Kochenderfer, M.J.; Gerdes, J.C. Value Sensitive Design for Autonomous Vehicle Motion Planning. In Proceedings of the 2018 IEEE Intelligent Vehicles Symposium (IV), Changshu, China, 26–30 June 2018; pp. 1157–1162.
32. Dadgar, M.; Joshi, K.D. The Role of Information and Communication Technology in Self-Management of Chronic Diseases: An Empirical Investigation through Value Sensitive Design. *J. Assoc. Inf. Syst.* **2018**, *19*, 86–112. [CrossRef]
33. Dechesne, F.; Warnier, M.; van den Hoven, J. Ethical Requirements for Reconfigurable Sensor Technology: A Challenge for Value Sensitive Design. *Ethics Inf. Technol.* **2013**, *15*, 173–181. [CrossRef]
34. Warnier, M.; Dechesne, F.; Brazier, F. Design for the Value of Privacy. In *Handbook of Ethics, Values, and Technological Design: Sources, Theory, Values and Application Domains*; van den Hoven, J., Vermaas, P.E., van de Poel, I., Eds.; Springer: Dordrecht, The Netherlands, 2014; pp. 1–14.
35. Friedman, B. *Human Values and the Design of Computer Technology*; Friedman, B., Ed.; CSLI Publications: Stanford, CA, USA, 1997.
36. Huldtgren, A. Design for Values in ICT. In *Handbook of Ethics, Values, and Technological Design: Sources, Theory, Values and Application Domains*; van den Hoven, J., Vermaas, P.E., van de Poel, I., Eds.; Springer: Dordrecht, The Netherlands, 2014; pp. 1–24.
37. Van den Hoven, J. ICT and Value Sensitive Design. In *The Information Society: Innovation, Legitimacy, Ethics and Democracy In honor of Professor Jacques Berleur s.j.: Proceedings of the Conference "Information Society: Governance, Ethics and Social Consequences", University of Namur, Namur, Belgium, 22–23 May 20*; Goujon, P., Lavelle, S., Duquenoy, P., Kimppa, K., Laurent, V., Eds.; Springer: Boston, MA, USA, 2007; pp. 67–72. [CrossRef]
38. Weibert, A.; Randall, D.; Wulf, V. Extending Value Sensitive Design to Off-the-Shelf Technology: Lessons Learned from a Local Intercultural Computer Club. *Interact. Comput.* **2017**, *29*, 715–736. [CrossRef]

39. Mueller, M.; Heger, O.; Niehaves, B. Exploring Ethical Design Dimensions of a Physiotherapeutic MHealth Solution through Value Sensitive Design. In Proceedings of the Hawaii International Conference on System Sciences (HICSS), Maui, HI, USA, 16 August 2018.
40. Mueller, M.; Heger, O. Health at Any Cost? Investigating Ethical Dimensions and Potential Conflicts of an Ambulatory Therapeutic Assistance System through Value Sensitive Design. In Proceedings of the Thirty Ninth International Conference on Information Systems, San Francisco, CA, USA, 13–16 December 2018.
41. Van Andel, J.; Leijten, F.; Van Delden, H.; van Thiel, G. What Makes a Good Home-Based Nocturnal Seizure Detector? A Value Sensitive Design. *PLoS ONE* **2015**, *10*, e0121446. [CrossRef] [PubMed]
42. Cheon, E.; Su, N.M. Integrating Roboticist Values into a Value Sensitive Design Framework for Humanoid Robots. In Proceedings of the Eleventh ACM/IEEE International Conference on Human Robot Interaction, Christchurch, New Zealand, 7–10 March 2016; IEEE Press: Piscataway, NJ, USA, 2016; pp. 375–382.
43. Timmermans, J.; Zhao, Y.; van den Hoven, J. Ethics and Nanopharmacy: Value Sensitive Design of New Drugs. *Nanoethics* **2011**, *5*, 269–283. [CrossRef] [PubMed]
44. van den Hoven, J. Nanotechnology and Privacy: The Instructive Case of RFID. In *Ethics and Emerging Technologies*; Sandler, R.L., Ed.; Palgrave Macmillan: London, UK, 2014; pp. 285–299.
45. Atomically Precise Manufacturing and Responsible Innovation: A Value Sensitive Design Approach to Explorative Nanophilosophy. Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3141478 (accessed on 02 January 2019).
46. Friedman, B.; Howe, D.C.; Felten, E. Informed Consent in the Mozilla Browser: Implementing Value-Sensitive Design. In Proceedings of the 35th Annual Hawaii International Conference on System Sciences, Big Island, HI, USA, 10 January 2002.
47. Doorn, N. Governance Experiments in Water Management: From Interests to Building Blocks. *Sci. Eng. Ethics* **2016**, *22*, 755–774. [CrossRef] [PubMed]
48. Van de Poel, I. *Translating Values into Design Requirements BT—Philosophy and Engineering: Reflections on Practice, Principles and Process*; Michelfelder, D.P., McCarthy, N., Goldberg, D.E., Eds.; Springer: Dordrecht, The Netherlands, 2013; pp. 253–266.
49. Richardson, H.S. *Practical Reasoning about Final Ends*; Cambridge University Press: Cambridge, UK, 1997.
50. Vermaas, P.E.; Hekkert, P.; Manders-Huits, N.; Tromp, N. Design Methods in Design for Values. In *Handbook of Ethics, Values, and Technological Design: Sources, Theory, Values and Application Domains*; van den Hoven, J., Vermaas, P.E., van de Poel, I., Eds.; Springer: Dordrecht, The Netherlands, 2014; pp. 1–19.
51. Baum, S.D. On the Promotion of Safe and Socially Beneficial Artificial Intelligence. *AI Soc.* **2017**, *32*, 543–551. [CrossRef]
52. Liaison Committee. *New Investigative Committees in the 2017–18 Session*; Authority of the House of Lords, The Stationery Office Limited: London, UK, 2017.
53. Lord Select Committee. Select Committee on Artificial Intelligence—Timeline. UK Parliament. Available online: https://www.parliament.uk/business/committees/committees-a-z/lords-select/ai-committee/timeline/ (accessed on 15 December 2018).
54. Franssen, M. Design for Values and Operator Roles in Sociotechnical Systems. In *Handbook of Ethics, Values, and Technological Design: Sources, Theory, Values and Application Domains*; Springer: Berlin/Heidelberg, Germany, 2015; pp. 117–149.
55. Friedman, B.; Kahn, P.H., Jr. Human Values, Ethics, and Design. In *The Human-Computer Interaction Handbook*; CRC Press: Boca Raton, FL, USA, 2007; pp. 1223–1248.
56. Lord Select Committee. *Select Committee on Artificial Intelligence. Collected Written Evidence Volume*; Lord Select Committee: London, UK, 2018.
57. Lord Select Committee. *Select Committee on Artificial Intelligence Collated Oral Evidence Volume*; Lord Select Committee: London, UK, 2018.
58. Secretary of State for Business, Energy and Industrial Strategy. *Government Response to House of Lords Artificial Intelligence Select Committee's Report on AI in the UK: Ready, Willing and Able?* Secretary of State for Business, Energy and Industrial Strategy: London, UK, 2018.
59. Select Committee on Artificial Intelligence. *AI in the UK: Ready, Willing and Able?* Select Committee on Artificial Intelligence: London, UK, 2018.

60. Hoff, R.D. Google Tries to Make Machine Learning a Little More Human. MIT Technology Review. Available online: https://www.technologyreview.com/s/542986/google-tries-to-make-machine-learning-a-little-more-human/ (accessed on 18 December 2018).
61. Harman, G. *Object-Oriented Ontology: A New Theory of Everything*; Penguin Random House: New York, NY, USA, 2018.
62. Morton, T. *Being Ecological*; MIT Press: Boston, MA, USA, 2018.